SPECIFICATION

NUMERICAL EXPRESSION RETRIEVING DEVICE

FIELD OF THE INVENTION

本発明は、自然言語で数量表現を検索する数量表現検索装置に関する。

BACKGROUND OF THE INVENTION

自然言語で様々に表記された、実質的に同じ意味の数量表現を検索できるようにするためには、数量表現を変換する必要がある。

例えば特開平 05 - 67137 号公報に記載されている、従来の数量表現検索装置では、文書内をサーチし、数量表現テンプレートとマッチングを行うことによって、文書内の数量表現を一括して適切な数量表現に変換することが可能となり、機械翻訳システムなどに利用することができる。

しかしながら、上記従来の数量表現検索装置では、単語の意味情報と変換関数を用いて変換するだけなので、省略されていて意味が複数考えられる表現には正確に対応することができない。

例えば従来技術では、昔の日本の長さの単位である「尺」(１尺は約１フィ

ート）を日本語の長さの数量表現に登録し、「centimeter」を英語の長さの数量表現に登録しておくと、「尺」を「centimeter」に変換することができると説明されているが、文書内で「キロ」と省略されている場合は、「キロメートル」なのか、「キログラム」なのか、判断できないので、正しく変換することはできなかった。

本発明は、このような従来の検索装置の課題に鑑みてなされたものであり、接頭語のみに省略されている場合を意識しなくても数量表現を検索することができる数量表現検索装置を提供することを目的とする。

SUMMARYOF THE INVENTION

上記問題点を解決するために、本発明の数量表現検索装置は、検索対象の文書や検索したい数量表現を入力する入力手段，上記入力された文書や数量表現の文の構造を解析する文構造解析手段，属性を表す属性名と属性の意味を表す属性内容並びに補完のための基本単位とからなる属性情報、および省略されたと判断するための接頭語と接頭語の意味を表す倍数とからなる単位系情報が格納された属性辞書，属性を表す属性名と属性名を判断するための共起語とからなる情報が格納された共起語辞書，および上記入力された文書や数量表現に対

して、上記解析された文構造と上記属性辞書を参照して、又は更に上記共起語辞書も参照して、上記文書や数量表現の接頭語に基本単位を補完することによって、省略された数量表現を補完する省略補完手段とを備える。

BRIEF DESCRIPTION OF THE DRAWINGS

図１は、本発明の実施例の数量表現検索装置のブロック構成図である。

図２は、数量表現を含む日本文の文構造の解析例を示す図である。

図３は、図１の属性辞書の構成例を示す図である。

図４は、図１の共起語辞書の構成例を示す図である。

図５は、図１の数量表現検索装置の動作を説明するフローチャートである。

図６は、図５のステップ 502 の投入処理の動作を説明するフローチャートである。

図７は、図６のステップ 602 における文構造の解析例を示す図である。

図８は、図５のステップ 503 の検索処理の動作を説明するフローチャートである。

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS OF

THE INVENTION

図１は、本発明の実施例の、数量表現検索装置のブロック構成図である。この実施例の数量表現検索装置は、入力手段１，分構造解解析手段２，省略補完手段３，属性辞書４，共起語辞書５，文書格納検索手段６，文書データベース７，抽出手段，及び出力手段９を備える。

入力手段１は、検索対象の文書や検索したい数量表現を入力する手段である。この入力手段１は、入力された文書や数量表現を分構造解析手段２に送る。

分構造解析手段２は、入力された文の構造を解析する手段である。この文構造解析手段２は、入力手段１から送られた文書や数量表現の文構造を形態素解析と構文解析によって解析し、解析後の文構造を入力された元の文書や数量表現とともに省略補完手段３に送る。

省略補完手段３は、接頭語のみに省略された（：省略されて接頭語のみ記載された）数量表現に基本単位を補完する手段である。この省略補完手段３は、分構造解析手段２から送られた文構造をもとに、属性辞書４と共起語辞書５を参照して、文書や数量表現の接頭語に基本単位を補完し、補完後の文書や数量表現を入力された元の文書や数量表現とともに抽出手段８に送る。

図２は数量表現を含む文の文構造の解析例を示す図である。なお、この実施例は日本語の文書に対する処理について説明されたものであり、英訳に際しては、必要により日本語の文書 or 文とそれに対応させた英語の文書 or 文を併記する。

数量表現の修飾先の語を、その共起語とする。図２（１），（２），および（３）の「５M」の共起語は「メモリ」になる。また図２（４）の「５M」の共起語は「増設する」になる。

属性辞書４は、属性の情報と単位系の情報を格納する辞書である。この属性辞書４は、属性情報が属性名と属性内容と基本単位からなり、単位系情報が接頭語と倍数と基本単位からなる。

共起語辞書５は、省略を補完する共起語の情報を格納する辞書である。この共起語辞書５は、属性名と共起語からなる。

図３は属性辞書４の構成例を示す図であり、図４は共起語辞書５の構成例を示す図である。

文書格納検索手段６は、文書を格納および検索する手段である。この文書格納検索手段６は、抽出手段８から入力された補完後の文書や元の文書や検索キ

ーワードを文書データベース７に格納し、抽出手段８から入力された補完後の数量表現が検索キーワードと一致する文書を文書データベース７から検索して、出力手段９に送る。

文書データベース７は、検索対象の文書や補完後の文書を格納するデータベースである。

抽出手段８は、検索キーワードを抽出する手段である。この抽出手段８は、省略補完手段３から入力された補完後の文書や数量表現と、補完後の語を抽出した検索キーワードを、文書格納検索手段６に送る。

出力手段９は、結果を出力する手段である。この出力手段９は、文書格納検索手段６から送られた検索結果を出力する。

なお、形態素解析を行う処理，構文解析を行う処理，文書をデータベース化する処理，文書を格納または検索する処理，および該当部分（：検索キーワード）を抽出する処理は、一般的な部分に関しては、既知の自然言語処理技術を用いて実行することができる。

図５は、本発明の実施例の、数量表現検索装置の動作を説明するフローチャートである。この図５において、入力手段１で処理を選択し（ステップ５０１）、

投入処理を実行するか（ステップ５０２）、検索処理を実行するか（ステップ５０３）、または終了する。

図６は、図５中のステップ５０２の、投入処理の動作を説明するフローチャートである。

図６の投入処理では、まず最初に、検索対象の文書を入力手段１に投入する（ステップ６０１）。

例えば、以下の例文（ａ）または（ｂ）を投入する。

「１０キロの荷物を持って歩いた」･･･（ａ）

「１０キロ、荷物を持って歩いた」･･･（ｂ）

入力手段１に投入された文書は、文書解析手段２に送られる。

次に、文構造解析手段２において、投入された文書の文構造を解析する（ステップ６０２）。

図７の（１）及び（２）は、例文（ａ）及び（ｂ）各々の文構造の解析例である。

文構造解析手段２での解析後の文構造は、入力手段１から送られた元の文書とともに、省略補完手段３に送られる。

次に、省略補完手段３において、解析後の文構造および属性辞書４（構成については、図３参照）の単位系情報を参照して、文書から接頭語を探す（ステップ６０３）。

例文（ａ）及び（ｂ）では、接頭語として、共に「キロ」が探される。

なお、上記ステップ６０３から以下のステップ６０７までの処理は、省略補完手段３においての処理である。

次に、文構造解析手段２で解析された文構造から共起語を決定する（ステップ６０４）。

例文（ａ）の共起語は「荷物」と決定される。

例文（ｂ）の共起語は「歩く」と決定される。

次に、共起語辞書５（構成については、図４参照）を参照して、属性（：属性名）を決定する（ステップ６０５）。

例文（ａ）の属性名は「WEIGHT」と決定される。

例文（ｂ）の属性名は「LENGTH」と決定される。

さらに、属性辞書４を参照して、基本単位を決定する（ステップ６０６）。

例文（ａ）では、属性が「WEIGHT」なので、基本単位は「グラム」と決定

される。

例文（ｂ）では、属性が「LENGTH」なので、基本単位は「メートル」と決定される。

そして、接頭語に基本単位を補完する（ステップ６０７）。

例文（ａ）では、接頭語「キロ」に基本単位「グラム」が補完され、「１０キログラムの荷物を持って歩いた」となる。

例文（ｂ）では、接頭語「キロ」に基本単位「メートル」が補完され。「１０キロメートル、荷物を持って歩いた」となる。

補完後の文書は、元の文書とともに、抽出手段８に送られる。

次に、抽出手段８において、補完後の語を検索キーワードとして抽出する(ステップ６０８）。

例文（ａ）では、「１０キログラム」がキーワードとして抽出される。

例文（ｂ）では、「１０キロメートル」がキーワードとして抽出される。

抽出されたキーワードは、元の文書とともに、文書格納検索手段６に送られる。

最後に、文書格納検索手段６によって、文書データベース７に、元の文書と

検索キーワードを格納し（ステップ６０９）、投入処理を終了する。

例文（ａ）では、元の文書「１０キロの荷物を持って歩いた」とキーワード「１０キログラム」が文書データベース７に格納される。

例文（ｂ）では、元の文書「１０キロ、荷物を持って歩いた」とキーワード「１０キロメートル」が文書データベース７に格納される。

図８は図５のステップ５０３の検索処理の動作を説明するフローチャートである。

図８の検索処理では、まず最初に、検索したい数量表現を、検索語として入力手段１に入力する（ステップ８０１）。

例えば、以下の例文（ｃ）または（ｄ）を検索語として入力する。

「１０キロメートル」･･･（ｃ）

「１０キロ」･･･（ｄ）

入力手段１に入力された数量表現（：検索語）は、文構造解析手段２に送られる。

次に、文構造解析手段２において、検索語の文構造を解析する（ステップ８０２）。文構造解析手段２での解析後の文構造は、入力手段１から送られた数量

表現（：検索語）とともに、省略補完手段３に送られる。

次に、省略補完手段３において、解析後の文構造および属性辞書４の単位系情報を参照して、検索語が接頭語であるか否か（接頭語のみに省略された数量表現であるか否か）を判別する（ステップ８０３）。

例文（ｃ）では、接頭語ではないと判断される。

例文（ｄ）では、「キロ」が接頭語であると判断される。

接頭語ではないと判別された場合は、文書格納検索手段６に送られる。

上記ステップ８０３で接頭語でないと判別された場合は、文書格納検索手段６において、検索後と検索キーワードが一致する文書を文書データベース７に格納されている文書から検索して獲得する（ステップ８０４）。

例文（ｃ）では、検索キーワードが「１０キロメートル」である上記例文（ｂ）の「１０キロ、荷物を持って歩いた」が、文書データベース７から検索されて獲得される。

そして、上記ステップ８０４で獲得した文書を検索結果として出力手段８から出力する（ステップ８０５）。

例文（ｃ）では、上記例文（ｂ）の「１０キロ、荷物を持って歩いた」が検

索結果として出力される。

また、上記ステップ８０３で接頭語であると判別された場合は、省略補完手段３において、属性辞書４の属性情報を参照して、基本単位と属性内容の一覧を出力手段９で表示させて、省略された数量表現であることを通知する（ステップ８１１）。

なお、上記ステップ８１１から以下のステップ８１５までの処理は、省略補完手段３においての処理である。

そして、再入力するか否かを出力手段９で表示させ、再入力するか否かを問い合わせる（ステップ８１２）。

上記ステップ８１２で、再入力しないことが選択された場合には、基本単位から選択するか否かを出力手段９で表示させ、基本単位から選択するか否かを問い合わせる（ステップ８１３）。

上記ステップ８１３で、基本単位からの選択がなされた場合には、その選択された基本単位で接頭語（：検索語）を補完する（ステップ８１４）。

例文（ｄ）では、例えば基本単位「グラム」が選択され、その基本単位「グラム」で検索語「１０キロ」が補完され、「１０キログラム」となる。

補完された検索語は、文書格納検索手段６に送られる。

そして、上記ステップ８１４で検索語に補完がなされた場合には、文書格納検索手段６において、検索語と検索キーワードが一致する文書を文書データベース７に格納されている文書から検索して獲得し（ステップ８０４）、この獲得した文書を検索結果として出力手段８から出力する（ステップ８０５）。

例文（ｄ）では、検索キーワードが「１０キログラム」である上記例文（ａ）の「１０キロの荷物を持って歩いた」が、文書データベース７から検索されて獲得され、検索結果として出力手段８から出力される。

また、上記ステップ８１３で、基本単位からの選択がなされなかった場合には、全ての基本単位で接頭語（：検索語）を補完する（ステップ８１５）。

例文(ｄ)では、例えば全ての基本単位「メートル」,「グラム」,「バイト」,･･･で、検索語「１０キロ」が補完され、「１０キロメートル」,「１０キログラム」,「１０キロバイト」,･･･となる。

全ての基本単位で補完されたこれら検索語は、文書格納検索手段６に送られる。

そして、上記検索ステップ８１５で検索語に補完がなされた場合には、文書

格納検索手段６において、補完された全ての検索語について、変換後の検索キーワードが一致する文書を文書データベース７に格納されている文書から、それぞれ検索して獲得し（ステップ８０４）、これら獲得した文書を検索結果として出力手段８から出力する（ステップ８０５）。

例文（ｄ）では、検索キーワードが「１０キログラム」である上記例文（ａ）の「１０キロの荷物を持って歩いた」や、検索キーワードが「１０キロメートル」である上記例文（ｂ）の「１０キロ、荷物を持って歩いた」などの例文が、文書データベース７から検索されて獲得され、検索結果として出力手段８から出力される。

共起語として修飾関係や格関係を有する語を抽出する方法や、抽出した語の関係を表すシソーラスを作成する手法や、抽出した語の関係から訳し別けを行う手法は既に存在するが、それらの手法の対象は、被修飾語や格になる名詞や動詞であり、それらの手法を用いても、修飾語になる数量表現の属性を決定することはできない。

以上のように、本発明の実施例によれば、文構造を解析して共起語を決定し、省略された数量表現を補完して格納しておく、あるいは省略された数量表現を

適切に補完する語のみを検索時に提供して検索することによって、省略された数量表現が検索対象文書と検索語のいずれに存在していても、数量表現検索装置が自動的に省略を補完して検索するので、利用者は、これらの省略を意識せずに検索を行うことができる。

また、自然言語の検索に用いると、数量表現の検索が容易になる。

なお、上記実施例では、数値と単位による数量表現のみを検索対象または検索語とする数量表現検索装置を説明したが、本発明は、その他の数量表現や数量表現以外の検索対象または検索語とする検索方法や装置と組合せて利用することも可能である。

また、上記実施例では、日本語からなる例文を用いて処理の詳細を説明しているが、本発明は日本語以外の言語、例えば英語や中国語であっても適用することができる。

さらに、上記実施例では、属性辞書に格納される基本単位として日本で一般的に用いられている単位である「メートル」や「グラム」を採用しているが、米国等で一般的に用いられている単位である「フィート」や「ポンド」を基本単位として採用することも可能である。

以上説明したように、本発明によれば、接頭語にのみ省略されていることを意識せずに数量表現を補完して検索できるという効果が得られる。

What is claimed is:

1. 自然言語で数量表現を検索する数量表現検索装置において、

検索対象の文書や検索したい数量表現を入力する入力手段と、

上記入力された文書や数量表現の文の構造を解析する文構造解析手段と、

属性を表す属性名と属性の意味を表す属性内容と補完するための基本単位とからなる属性情報、および省略されたと判断するための接頭語と接頭語の意味を表す倍数とからなる単位系情報が格納された属性辞書と、

属性を表す属性名と属性名を判断するための共起語とからなる情報が格納された共起語辞書と、

上記入力された文書や数量表現に対して、上記解析された文構造と上記属性辞書を参照して、又は更に上記共起語辞書も参照して、上記文書や数量表現の接頭語に基本単位を補完することによって、省略された数量表現を補完する省略補完手段と

を備えたことを特徴とする数量表現検索装置。

2. 請求項１記載の数量表現検索装置において、

上記保管後の文書から接頭語に基本単位を補完した語を検索キーワードと

して抽出する抽出手段と、

文書データが格納される文書データベースと、

上記補完後の文書と上記入力された元の文書と上記抽出された検索キーワードとを上記文書データベースに格納する文書格納検索手段と

を更に備え、

上記省略補完手段は、上記解析された文構造と上記共起語辞書を参照して上記入力された文書から接頭語のみに省略された数量表現を探し、その省略された数量表現に対して、上記解析された文構造をもとに上記接頭語の共起語を決定し、その決定した共起語をもとに上記共起語辞書を参照して上記接頭語の属性名を決定し、この決定した属性名をもとに上記属性辞書を参照して上記接頭語に基本単位を補完する

ことを特徴とする数量表現検索装置。

3. 請求項２記載の数量表現検索装置において、

出力手段を更に備え、

上記省略補完手段は、上記解析された文構造と上記共起語辞書を参照して上記入力された数量表現が接頭語のみに省略された数量表現であるか否かを判別

し、接頭語のみに省略された数量表現であれば、その旨を上記出力手段によって通知し、最入力を促す

ことを特徴とする数量表現検索装置。

4. 請求項２記載の数量表現検索装置において、

出力手段を更に備え、

上記省略補完手段は、上記解析された文構造と上記共起語辞書を参照して上記入力された数量表現が接頭語のみに省略された数量表現であるか否かを判別し、接頭語のみに省略された数量表現であれば、基本単位と属性情報を上記出力手段によって提示して、基本単位の内の１つを選択するように促し、選択された基本単位を用いて補完する

ことを特徴とする数量表現検索装置。

5. 請求項４記載の数量表現検索装置において、

上記省略補完手段は、省略された数量表現を補完するための基本単位が選択されなかったときに、補完可能な全ての基本単位を用いて補完する

ことを特徴とする数量表現検索装置。

6. 請求項３記載の数量表現検索装置において、

上記文書格納検索手段は、上記入力された数量表現と検索キーワードが一致する文書を上記文書データベースから検索し、上記検索結果として上記出力手段によって出力する

ことを特徴とする数量表現検索装置。

ABSTRACT

接頭語のみに省略されている場合を意識しなくても、数量表現を検索することができる、数量表現検索装置を実現するために、本発明の数量表現検索装置は、検索対象の文書や検索したい数量表現を入力する入力手段１,上記入力された文書や数量表現の文の構造を解析する文構造解析手段２,属性を表す属性名と属性の意味を表す属性内容並びに補完のための基本単位とからなる属性情報、および省略されたと判断するための接頭語と接頭語の意味を表す倍数とからなる単位系情報が格納された属性辞書４,属性を表す属性名と属性名を判断するための共起語とからなる情報が格納された共起語辞書５,および上記入力された文書や数量表現に対して、上記解析された文構造と属性辞書 4 を参照して、又は更に共起語辞書 5 も参照して、上記文書や数量表現の接頭語に基本単位を補完することによって、省略された数量表現を補完する省略補完手段３とを備える。